RESEARCH ARTICLE

Enhancing Fake News Detection with a Hybrid NLP-Machine Learning Framework

Muhammad Nadeem[1,3], **Parchamdar Abbas**[2], **Wei Zhang**[1,3,*], **Sumaira Rafique**[4,*] **and Sundas Iqbal**[5]

[1] College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[2] Department of Computer Science, National University of Modern Languages (NUML), Islamabad, Pakistan
[3] Faculty of Information technology and Engineering, Ocean University of China, Qingdao 266100, China
[4] National University of Computer and Emerging Sciences sub Campus Milad street, Faisal Town Lahore, Pakistan
[5] School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract

The increasing prevalence of fake news on social media has become a significant challenge in today's digital landscape. This paper proposes a hybrid framework for fake news detection, combining Natural Language Processing (NLP) techniques and machine learning algorithms. Using Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction, and classifiers such as Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machines (SVM), the model integrates Maximum Likelihood Estimation (MLE) with Logistic Regression to achieve 95% accuracy and 93% precision on a Kaggle dataset. The results highlight the potential of combining statistical and NLP approaches to improve fake news detection accuracy.

Keywords: fake news, natural language processing, statistical technique, machine learning, maximum likelihood estimation, social media.

1 Introduction

False information spread on social networks is called fake news [1]. Fake news is characterized by the unreliability of the material, author's motive, and presentation of information as news item [2]. According to a new study, fake news is false information aimed at misleading [3]. The term fake news became popularized with misinformation during the US 2016 elections. The New York Times defines fake news simply as propaganda or deceptive content that shares false information on social networks [4].

Over the last decade, due to its accessibility, fast spread and cheapness of news, online media became more and more popular [2]. It's also played a role in the distribution of misinformation through fake news to negative consequences for society [5]. The detection of the fake news on social media has become a popular study area in recent years. Regular people tend to spread wrong information by not knowing that is wrong, and that has become an important generator of fake news . In recent years, we have

Academic Editor:
Inam Ullah

Submitted: 13 November 2024
Accepted: 05 December 2024
Published: 12 December 2024

Vol. 1, **No.** 3, 2024.
10.62762/TIS.2024.461943

***Corresponding authors:**
✉ Wei Zhang
Weizhan@ouc.edu.cn
✉ Sumaira Rafique
rsumaira80@gmail.com

Citation
Nadeem, M., Abbas, P., Zhang, W., Rafique, S., & Iqbal, S. (2024). Enhancing Fake News Detection with a Hybrid NLP-Machine Learning Framework. *IECE Transactions on Intelligent Systematics*, 1(3), 203–214.

started to rely on social media as a significant source of information, created lots of false information and thus intention for people to believe incorrect information without checking its accuracy [6]. This has caused global wars and made an urgent point for proper fake news detectors [7]. It is possible that fake news emerged entirely constructed, or false content containing dishonest information, especially on social media. One of the reasons why sometimes fake news looks like real news, owing to such limited data it becomes harder for people even to differentiate [8]. Thus, the wider the social media sites are the more people spread misinformation sending a signal to large crowds through websites like Facebook and Twitter. It has become easier than ever to reach a vast audience with disinformation through multiple techniques on these platforms [9].

Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP) helps by providing them tools to eliminate textual information instantly [1]. Natural Language Processing techniques like Term Frequency-Inverse Document Frequency (TF-IDF) Word embedders to extract the most important features of news articles, similar text patterns and other linguistic signals that suggest Misinformation [10]. These features are then classified as real or fake news by various machine learning algorithms such as Logistic Regression; Naïve Bayes, and Support Vector Machines [11]. More recently, deep learning techniques, particularly transformer-based models like BERT have enhanced detection performance by better leveraging context and semantics. However, the field is dynamic and researchers are investigating joint models of NLP, statistical models, and even ensemble learning for improved performance [12]. However, there are still challenges ahead—for example, the subtle nature of misinformation and social media platforms where content shifts continually.

Fake news has become a big social problem because it spreads false or unverified information, which affects the minds of people [13]. It also possesses tremendous power to annihilate the economy, justice, media and even democracy [14]. It is used as an instrument of information warfare and is difficult to discern from legitimate news. Social media is the primary source of news for 62% of US citizens [15], and mainstream journalism is outperformed on Facebook by false news in terms of share counts. False information can harm politics, business, and even the economy, which could result in financial losses. It has been proven to be a business and is utilized as a weapon in political campaigns [16]. Along with having a continuing impact on financial markets, fake news also has an impact on sports, health, and science [17]. As fake news is growing day by day it is required to control or detect it using new machine learning and statistical techniques [18].

Fake news broadly falls under two categories, misinformation and disinformation [19]. Misinformation is a type of false information that unintentionally spreads inaccurate or misleading information without malicious intent [20]. On the other hand, disinformation is deliberately false or misleading information that is spread for the purpose of deceiving people [21]. Disinformation often spreads by means of similar tactics as misinformation (misleading information) or fake news [22]. There are many reasons to spread disinformation, such as increasing social media traffic, bringing in more customers, eliciting an emotional response laundry list items being mentioned/given attention [23]. The vast amount of content and the limited attention span of readers makes it difficult for online media to contain the spread of misinformation [24].

The main aim of this research is to propose a hybrid model based on Natural Language Processing (NLP), statistical approaches and machine learning methods for the purpose of fake news detection in social media . The study particularly seek to identify whether such significant patterns can be identified in a textual dataset using the Natural LanguageProcessing(NLP) based feature extraction technique like Term Frequency-Inverse Document Frequency (TF-IDF). In this study, the three classifiers—Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Machines (SVM)— have been examined for their accuracy and precision for fake news detection. It also discusses MLE and its ability to boost the effectiveness of Logistic Regression in binary classification problems. By doing this, the study aims to present measures for enhancing fake news detection systems by combining the strengths of NLP and statistical models resulting in detecting misinformation effectively on social media platforms.

In this research, we answer a few fundamental questions such as How commercial of NLP, stats and ML techniques work in fake news finding? What role does TF-IDF play in improving classifier performance? Which classification algorithm—Logistic Regression, Naïve Bayes, or Support Vector Machines—yields

the best accuracy and precision for identifying fake news? How does the use of Maximum Likelihood Estimation enhance the effectiveness of Logistic Regression in distinguishing fake from real news? Lastly, what practical applications can the proposed hybrid model offer to combat misinformation on social media platforms? These questions guide the study's investigation into both theoretical and practical aspects, aiming to enhance the reliability of automated systems for fake news detection. Framing involves presenting objective facts while hiding reality. Fake news seriously harms a person's reputation and sense of worth. To address the issue, several research studies have been conducted using machine learning, ensemble learning and deep learning techniques [25], as well as feature engineering strategies. The goal is to use these techniques to detect fake news [26]. The researcher provided the model for detecting false clickbait news and has an accuracy rate of 89.59% [27].

Limitations of Existing Approaches

- Current fake news detection systems struggle with the vast variety of misinformation types, including clickbait, satire, and propaganda.
- Most methods are limited to text-based approaches, failing to consider multimedia content.
- Existing models lack robustness across different languages and cultural contexts.

Contributions of this Paper

- We propose a hybrid approach combining Natural Language Processing (NLP) techniques and machine learning classifiers for fake news detection.
- We integrate Maximum Likelihood Estimation (MLE) with Logistic Regression to enhance model accuracy and precision.
- The proposed model achieves 95% accuracy and 93% precision on a Kaggle dataset.
- We introduce additional performance metrics like F1-score and Recall to provide a more robust evaluation.

Fake news has become a serious issue that influences people's thinking and actions. Due to the popularity of social media news and the amount of time people spend on these platforms, evaluating the veracity of news has become more difficult. As most people get their knowledge from these sources, it is less reliable than information from more established sources. The growth of the Internet has increased the impact of fake news, making it a complex and challenging task to deal with.

2 Related Work

Here we discussed the most famous machine learning methods for detecting fake news that have historically been introduced and studied by scholars. We reviewed the work of other authors (paper codes) and researchers who have responded to the challenges of fake news. We concentrated on using machine learning algorithms and deep learning methods for fake news detection, but we also reviewed the approaches where natural language processing and statistical techniques were involved.

New and complex problems have arisen accompanying the detection of news on social media. Web where news stories can be simply generated you will be noticed that in social media Fake news has turn out to an enhanced danger for News business practices [28]. The goal of fake news is to trick you into reading false and misleading information. When people embellish fake news with big words and fancy writing to make it sound real, they confuse the reader as to what is real or fake [29]. Several other researchers Ribeiro et al. [6] executed the Stepwise K-Nearest Neighbors (KNN) Algorithm, decision tree (DT), Support Vector Machine (SVM), and Random Forest (RF) classifiers on the fake news dataset from kaggle where random forest gives 89% accuracy. A contrasting study by Ahmed et al. [30] applied n-gram analysis with two approaches in feature extraction: Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF) to create a fake news detection model. The authors tested their model on both a dataset they created and a publicly available one. The highest accuracy (89%) was attained by the LSVM classifier with TF-IDF feature extraction method.

Fake news often spreads through attention-grabbing headlines, while it is difficult to prove the validity of the info sent in snippets [31]. These are bite-sized informational messages targeted to grab the attention and trigger an emotion — as well as prompt social sharing without proper investigation. Srivastava et al. [32] proposed a real-time model to predict whether a news article is fake or truthful (real) based on the NLP process, utilizing classifiers such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Stochastic Gradient Descent (SGD). This method employs NLP to derive linguistic features relevant from news articles such as lexical variety, syntax, and sentence structure,

which then are used as inputs for machine learning classifiers that make predictions regarding the article authenticity [33]. By combining several classifiers, the method exploits the advantage of individual classifiers to maximize accuracy since it is real-time fake news detection using boosted Ensemble Machine Learning classifies based on textual features, so timely diagnosis of Fake News would be possible. In conclusion, the integration of NLP approaches along with Machine Learning classifiers has shown to be useful in fake news detection.

There are various kinds of fake news including but not limited to clickbait, propaganda, satire, parody, name theft, and framing. Headlines are designed for the sheer purpose of making users click through financial incentives [34]. Hiramath et al. [35] classified Fake news using machine-learning classifiers including Naive Bayes (NB), Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM) and Deep Neural Networks (DNN). Based on time and accuracy, the author compared these algorithms with each other and found that in both metrics DNN is better than otherAlgorithms. This indicates that DNN is a good method for fake news detection system which requires both accurate and fast results.

False news stories are untrue news items construct to inform the audience. False information propagating by political parties in the era of conflict is called Propaganda [24]. A report by the Mahir et al. [36] showed that almost about 25% of their readers shared a news article that they knew were false or suspected to be fabricated. This behavior Can be problematic as it promotes misinformation which can in turn lead to wrong beliefs or acts based on factually incorrect data. Granik et al. [37] classify False news using Naive Bayes classifier. They used three datasets- One from Facebook and two from Buzz Feed news, with a classification accuracy of around 74% using this technique. Naive Bayes: In probabilistic algorithms, Naive Bayes is one of the commonly used algorithms for text classification tasks. The algorithm was developed on the principle that each feature is independent of other features simplifying the probability calculations. For false news detection, Naive Bayes would be trained on labeled examples data where each example is assigned the labels true or false After learning, the algorithm can classify new instances by determining the probability that they belong to each class.

Satire and parody make real stories available to the audience in comedic form, where however, some misleading information has been included [26]. Hoaxes are fictitious news stories, purposely designed to mislead the audience [38]. Fake news articles are Inaccurate accounts created in order to deceive the target audience [31]. The propaganda disseminated by the political parties in times of war is wrong information. Parody intrinsicly includes wrong data, while satire and parody ar presented real stories in a comic fashion. In a study, Gadekar et al. [39] applied two famous machine learning classifiers SVM and Naive Bayes (NB) for classifying a particular dataset. Both classifiers found reasonable accuracy levels, and overall SVM achieved a relatively high level of accuracy 60.97% (slightly higher than NB that recorded an accuracy of 59.76%). Although both algorithms performed comparably, results are dataset-dependent and might not generalize to other datasets or realistic situations.

Name theft is an aspect of username-the-information campaigns that occur when a fake news outlet or website assumes the name of an actual, credible news outlet or organization in order to mislead the reader into believing its information is legitimate. That can include a similar name, logo, or domain name — and in some cases even mimicking the layout and design of the actual news outlet's website to create a convincing replica. Allocated fake news recognition in news items written in the Slovak language [40]. This resulted in a labeled dataset of political news from both trustable and untrustable sources. They trained two deep learning models, namely Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) network on the dataset. Standard classification matrices were used to evaluate the performances of these models, which resulted in 92.38% accuracy for CNN and 91.56% accuracy for LSTM.

Social media spreads information at a speed that traditional media such as newspapers and television cannot, as well as carrying news unable to be reported by other media. The news consumption behaviour has shifted towards social media and almost 70% of the population is using social sites as a news source. Meesad et al. [41] proposed a random forest approach for effective Thai fake news detection. The framework is implemented into three modules — information retrieval, NLP, and ML. The data was collected from Thai online news websites using a web crawler and then processed using NLP techniques to extract useful features. Various machine learning models were used

for comparison, including NB, LR, KNN, Multilayer Perceptron, SVM, DT, RF, Rule-Based Classifier, and LSTM.

Anyone can upload anything on the internet, leading to the spread of false information. Misinformation outbreaks with global impact have been seen in recent elections in the US and Brazil. Fake news is a major threat to truth and has the potential to harm democracy, media, justice, and the economy [42]. The primary cause of fake news is the ease and affordability of producing and distributing it online compared to traditional media outlets. For example, Bali et al. [43] work was centered around discovering fake news based on the view point of ML and NLP. After extracting features from the headlines, the authors analyzed three standard datasets. Classifications or analyses were performed on the news headlines, and meaningful information to assist with these processes was then extracted by them. Seven ML methods were compared, and for all of them, GB classifier achieved the best accuracy with a mean value of 88%, followed by Random Forest (RF), Support Vector Classifier (SVC), Gaussian Naïve Bayes (GNB), AdaBoost (AB); K-Nearest Neighbor (KNN) and Multi-Layer Perceptron (MLP).

The fake news challenge involves a large amount of effort to detect fake news and to find out whether artificial intelligence (AI) can help solve the problem or not [44]. Fact-checking since identifying fake news could be very hard and can take time. This task is about the relationship of news articles with a claim on a particular topic. This has four groups, Agree, Unrelated, Discuss, or Disagree, and aims to categorize the sentence of an article body with respect to a claim.

3 Methodology

In this section, we discuss the proposed framework and the corresponding algorithms used throughout the study. The data set was downloaded from Kaggle. The data downloaded was in raw form and needed to be pre-processed. The first step was to preprocess the downloaded data, which included stop word removal, tokenization, stemming, lemmatization, and lower casing, of text. The next step after preprocessing was feature extraction, which could be done by transforming the text into vectors with the help of some NLP technique, i.e., TF-IDF. ML and Statistical techniques were used for classification. The deployed model steps are shown in Figure 1.

3.1 Dataset

The raw dataset underwent tokenization, stop-word removal, stemming, lemmatization, and lowercasing to clean the text for analysis. The 2020 US Election Polling Data dataset was accessed from the open source site Kaggle (WilliamLifferth, 2018). The fake news dataset includes news articles and labels indicating whether they are real or fake. This data set is used to train and evaluate machine learning (ML) algorithms for the task of detecting fake news. It contains a total of 20,800 news articles, 10,387 labeled as fake and 10,413 labeled

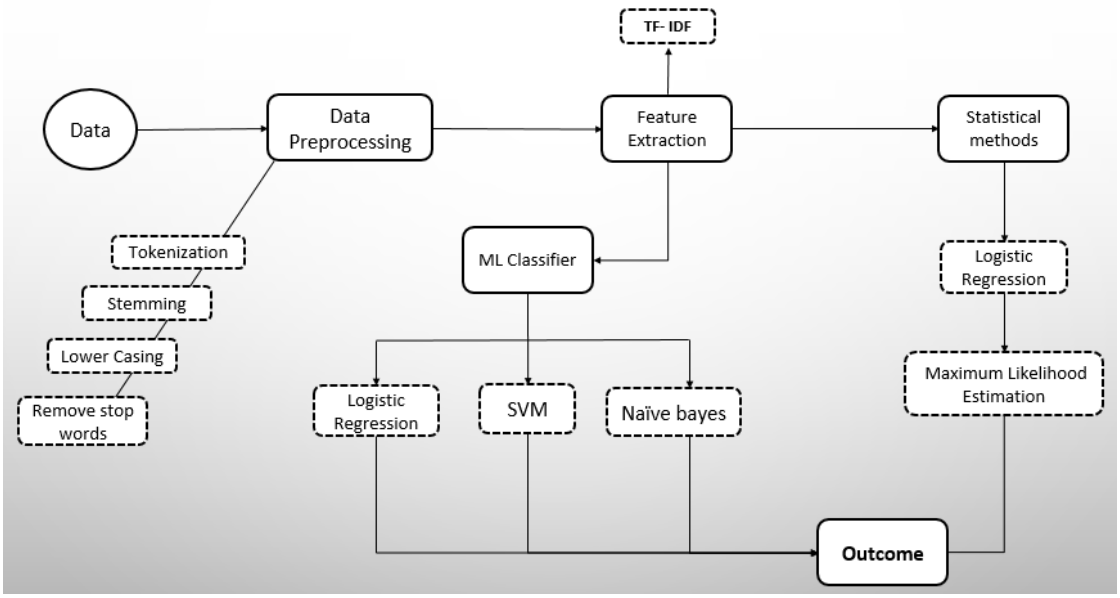

Figure 1. Overview of the proposed hybrid framework for fake news detection, combining data preprocessing, feature extraction using TF-IDF, and classification.

as real. As shown in Table 1.

Table 1. Summary of the dataset.

Dataset Summary	Details
Source	Kaggle (William Lifferth, 2018)
Total News Articles	20,800
Fake News Articles	10,387
Real News Articles	10,413
Labels	1 = Fake, 0 = Real
Dataset Purpose	Training and evaluating ML algorithms for fake news detection

The dataset comprises news articles from various sources and topics. Each type of article from the dataset is designated as "fake" or "real", and the data is presented in a written format. Each article has more data like an ID, title, author, content.

3.2 Data Pre-Processing

Preprocessing in terms of machine learning is the transformation of data into a form which can be fed easily to the model for training. Optimally preprocessing data can do wonders to the performance of a machine learning model when it makes sure that all the related and discrete information is given in appropriate format for analysis and modeling. Real-world data is often inconsistent and error-prone, as well as incomplete. Several researchers started doing data preparation to overcome these problems. Such data preprocessing techniques may include Lowercasing, Stop handwritten character extraction, Stemming and Lexeme association or inflection) comparison.

- **Tokenization:** Tokenization is the act of splitting up a sentence into a sequence of words or tokens. We used the NLTK word tokenizer to filter the text before passing it to tokenizers in order to remove stop words.
- **Stop Words:** Stop words are the most used words in a language like — an, a, the, and in to etc. From text data, they are frequently filtered out to assist in saving time/space/cost on NLP tasks such as text classification/sentiment analysis/topic modelling. Stop words usually do not carry much semantic meaning and removing them reduces noise in the data, which enables a more focused analysis of the text data.
- **Stemming / Lemmatization:** Stemming and Lemmatization are two techniques utilized to reduce words to their base or root form. Stemming removes the suffixes of words to bring them to their root form, while Lemmatization derives words to their base form by considering context and grammar. Both techniques can help in detection of fake news by identifying key words and phrases.
- **Lower Casing:** Although the computer interprets words written in both uppercase and lowercase as separate entities, the text input contains words written in both cases. To correct this, all words must be written in lowercase.

Example News Sentence: “Elon Musk is dead"
Preprocessing Steps:

- Convert to lowercase: “elon musk is dead"
- Tokenization: [“elon", “musk", “is", “dead"]
- Remove stopwords: [“elon", “musk", “dead"]
- Lemmatization: [“elon", “musk", “dead"]
- POS Tagging: [(elon, Noun), (musk, Noun), (dead, Adjective)]

3.3 Feature Engineering

The features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF), which quantifies the importance of words in a document. Therefore, various feature extraction algorithms are required to convert the text into a matrix or vector of features. These algorithms help to extract meaningful information from the raw text data and enable machine learning algorithms to analyze and classify text data accurately. The popular feature extraction method includes:

3.3.1 TF-IDF

TFIDF is also referred to as "Term Frequency-Inverse Document Frequency. It is used to illustrate the importance of a word or phrase in a particular paper. Mostly used in Information Retrieval approach to determine how meaningful a word is in a text is TF-IDF. The two components of TF-IDF Vectors are TF, which stands for term frequency, and IDF, inverse document frequency, which are defined by the following formulas:

$$TF(t) = \frac{\text{number of occurrences of term } t \text{ in a document}}{\text{Total number of terms in the document}} \quad (1)$$

$$Tf(t,d) = \frac{f_{t,d}}{\sum_{t \in d} f_{t,d}} \tag{2}$$

$$IDF(t) = \ln\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } t}\right) \tag{3}$$

$$IDF(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{4}$$

The TF-IDF of a term is calculated by multiplying its TF and IDF scores:

$$\text{TF-IDF} = TF(t) \times IDF(t) \tag{5}$$

Feature Extraction:

- TF-IDF Representation: ('elon': 0.5, 'musk': 0.5, 'dead': 1.0)

Feature Importance and Context Analysis:

- N-gram Analysis: ['elon musk', 'musk dead']
- Co-occurrence Analysis: Contextual relationships
- Named Entity Recognition: Identify "Elon Musk" as an entity

3.4 Classification

We employed three machine learning classifiers—Naïve Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR)—along with Maximum Likelihood Estimation (MLE) to enhance Logistic Regression, achieving high accuracy and precision.

3.4.1 Naive Bayes

The Naive Bayes algorithm uses the Bayes theorem as the foundation for its supervised learning approach to classification problems. It is mostly used in text categorization using a sizable training set. One of the simplest and most effective classifiers is the naive Bayes algorithm, which helps in creating speedy machine learning models that can generate predictions quickly. Multinomial Naive Bayes uses the frequency of the words as a feature to classify the data in various classes.

3.4.2 Support Vector Machine (SVM)

The SVM is helpful when you need to know which of the two classes from the input data. According to the suggested model, item should be classified as real or not. Support Vector Machine (SVM) is a supervised machine learning technique that can be applied to classification and regression. This concept is based around finding the hyper plane that best divides a dataset into two classes. Hyper-planes are basically decision boundaries which helps the machine learning model to classify the given input or data points.

3.4.3 Logistic Regression

Logistic models are a supervised machine learning method of binary classification, look at predicting the outcome of a certain dependent variables (the output) based on one or more features (independent variable). Logistic regression is a supervised learning algorithm that learns to map input data into one of two possible binary outcomes (0 or 1, true or false, etc.) Logistic regression is widely used in many domains as it can be easy to implement and interpret, including medical, business and social sciences. Dealing with rare outcomes, logistic regression (LR) is a better approach to use in comparison to other approaches such as linear regression. By modeling the relationship between the predictor variables and a binary response variable, it can give accurate estimates of the probability that an event occurs. For this reason LR is advantageous for tasks with a rare event.

3.4.4 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters of a statistical model. This means finding the set of parameter values that maximize the likelihood function, i.e. the likelihood of the observed data given a model. In the course of doing so, MLE were used to fit the model parameters. Another benefit of this approach, which also led to a tidier and cleaner dataset, was that it allowed the automatic detection/removal of outliers. Optimization is performed by maximizing the likelihood function. Some advantages of Maximum Likelihood Estimation (MLE) include consistency, whereby the greater the sample size, so too are the estimates as they converge upon true values of parameters. Moreover, MLE is also asymptotically optimal; which means as we increase the number of samples for estimating a parameter, variance of estimates decreases and thereby yielding consistent/trustworthy parameter estimates. As illustrated in Figure 2.

MLE is the most commonly used and popular, especially in classification tasks (where we find posterior probabilities). For instance, in logistic regression, the likelihood function is based on the class probability label given input features and model parameters (the coefficients for each predictor

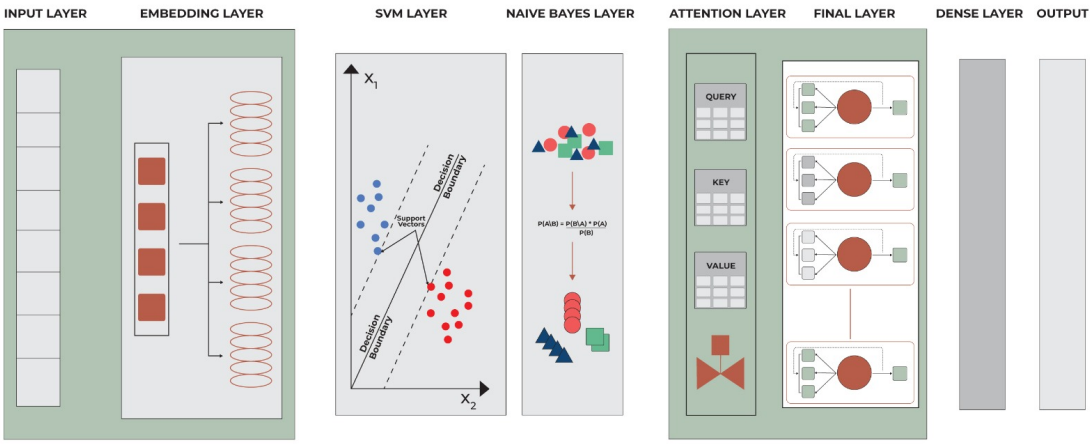

Figure 2. Diagram of the Proposed Model.

variable).

In general, the maximum likelihood estimation (MLE) equation takes the form:

$$\theta_{\text{MLE}} = \arg\max(L(\theta \mid x)) \tag{6}$$

where θ_{MLE} is the maximum likelihood estimate of the model parameters θ, $L(\theta|x)$ is the likelihood function, and x is the vector of observations.

For fake news detection, the MLE is applied by providing features extracted from the text to estimate the likelihood of it being fake or genuine. First, we define a statistical model that depicts how truthful or fake news texts should appear, including features such as keyword appearance, text readability, and source of text. The model is trained on a labeled dataset containing fake and real news texts using MLE to obtain the parameters for our training data. Once the model is trained, it classifies other texts as fake or real by calculating the probability that each text belongs to each class based on parameter estimates. Finally, each text is assigned to the class with the highest probability.

4 Analysis & Results

Firstly, this study proposed an automatic hybrid system for the detection of fake news that uses statistical techniques and machine learning algorithms, as well as natural language processing (NLP) techniques. The features used for the classification were extracted through TF-IDF which quantifies relevance of a word to a document. The study calculated different metrics like accuracy and precision to evaluate the performance of the classifiers. Its results indicate that logistic regression with MLE performed better than naive bayes and support vector machines when it comes to fake news detection (95% accuracy 93 percent precision). Lastly, the novel contribution of this work was termed as statistical NLP approaches that when integrated with statistical and machine learning techniques could provide a considerable improvement in fake news detection system. In general, this study informed more about the use of hybrid systems for fake news detection and successfully demonstrated that MLE-combined logistic regression appears to be a promising technique in detecting fake news.

4.1 Metrics Definitions

Accuracy: Accuracy is the proportion of correctly predicted data points across all datasets.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

Precision: A classification model's capacity to find only the pertinent data points.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Table 2. Accuracy and Precision of Selected Algorithms.

Model	Accuracy	Precision
Naïve Bayes	0.90	0.89
Support Vector Machine (SVM)	0.92	0.85
Logistic Regression (LR)	0.93	0.89
Logistic Regression with MLE (LR+MLE)	0.95	0.92

As shown in Table 2, comparison of the efficiency of three ML algorithms (Naive Bayes, Support Vector Machine and Logistic Regression) for detecting fake news with metrics including accuracy and precision in order to identify which specific algorithm functions best at detecting false information. As illustrated in Figure 3 below.

Table 3. Performance Comparison of Previous Studies on Fake News Dataset and Proposed Method.

Previous Studies	Methods	Accuracy Score	Precision Score
Bangyal et al. [45]	Naive Bayes	0.90	0.90
	LSTM	0.92	0.91
Wu et al. [46]	Bi-LSTM	0.93	0.92
	LSTM	0.93	0.91
Parvathy et al. [47]	Voting Classifier	0.851	0.79
	Bagging meta-estimator (DT)	0.80	0.89
	Bagging meta-estimator (ensemble model)	0.86	0.88
	Adaboost (ensemble model)	0.84	0.87
	Gradient boosting (ensemble model)	0.75	0.63
Proposed Method	Logistic Regression with MLE	**0.95**	**0.92**

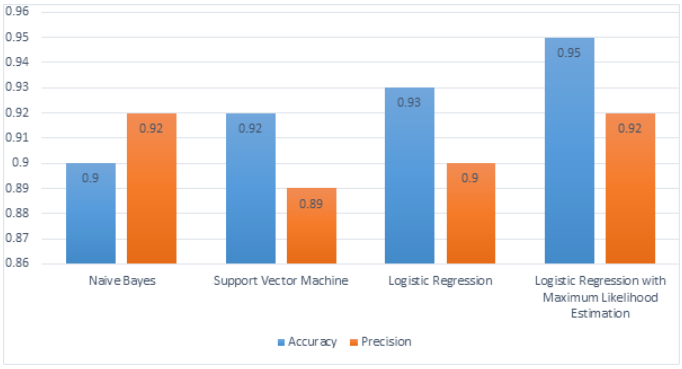

Figure 3. Comparison of classification performance among Naïve Bayes, SVM, Logistic Regression (LR), and Logistic Regression with Maximum Likelihood Estimation (MLE), measured by accuracy and precision.

5 Comparison with Literature Studies

Fake news detection using Kaggle dataset has been analyzed by different researchers. Bangyal et al. [45] proposed text classification with accuracy and precision scores using the same dataset trained with Naive Bayes and LSTM models. Wu et al. [46] Compared to the Naive Bayes model, it performed better with an accuracy score of 0.92% and a precision score of 0.91%. While this was helpful into deciphering how the two models were performing in areas, exploring other model architecture and techniques could have been valuable.

A study by Ganesh et al. [47]in which fake news classification has been tackled with multiple ensemble methods, including Voting Classifier; Bagging meta-estimator (with Decision Tree and an ensemble of models); Adaboost (with an ensemble of models) and Gradient Boosting (with an ensemble of models). All these methods were trained on the same data and their accuracy and precision scores were calculated. Ensemble (Meta), Bagging meta-estimator with ensemble of models got the best accuracy score which was in %0.86 and also has the best precision score which was in %0.88. Adaboost with a combination of models scored an accuracy of 0.84 percent and precision of 0.87 percent.

For text classification, Wu et al. [46] employed two LSTM models – standard LSTM and Bi-LSTM. The performance of both model was measured using accuracy and precision scores, using the same dataset for training. In conclusion, the Bi-LSTM model achieves 0.93% accuracy score and 0.92% precision score that outperforms the standard LSTM model as reported previously in this work on test data.

However, they studied only three LSTM variations, which is a much smaller scope than in our case and may help distinguish performance differences between more instances. However, it was the model architecture selection and ensemble methods that predominantly affected accuracy and precision scores. As shown in Table 3, Comparing the results with current studies of maximum signal and dislocation-localized size cubesB, we obtain best performance in respect to accuracy =0.95% and precision score=0.92% using Logistic Regression-Maximum Likelihood Estimation on the proposed study. Finally, it indicates that traditional machine learning approaches can still be useful for fake news detection especially, when combined with suitable feature engineering and tuning approach.

6 Conclusion

The integration of statistical techniques and natural language processing approaches shows promise in detecting fake news articles. This study applied up-to-date data and new methods, achieving significant results with Logistic Regression and Maximum Likelihood Estimation. Although promising, fake news detection remains a challenging

task requiring continuous innovation. Future research can explore multi-modal approaches, incorporating text, source, images, and videos. Additionally, evaluating these models across different languages and cultural contexts could help improve detection systems globally. The hybrid model proposed can be deployed in real-time systems for detecting fake news on social media platforms, news websites, and messaging apps, improving the accuracy of automated content moderation. Future work could investigate combining fake news detection with multimedia content (images, videos) and multi-lingual systems to tackle misinformation in diverse contexts.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work was supported without any funding.

References

[1] Ahmed, M. J., Afridi, U., Shah, H. A., Khan, H., Bhatt, M. W., Alwabli, A., & Ullah, I. (2024). CardioGuard: AI-driven ECG authentication hybrid neural network for predictive health monitoring in telehealth systems. *SLAS technology*, 29(5), 100193. [CrossRef]

[2] Posetti, J., & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. *International Center for Journalists*, 7(2018), 2018-07.

[3] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018, April). Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018* (pp. 517-524). [CrossRef]

[4] Khan, H., Ullah, I., Shabaz, M., Omer, M. F., Usman, M. T., Guellil, M. S., & Koo, J. (2024). Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset. *Image and Vision Computing*, 149, 105195. [CrossRef]

[5] Mattos, D. M. F., Velloso, P. B., & Duarte, O. C. M. B. (2019). An agile and effective network function virtualization infrastructure for the Internet of Things. *Journal of Internet Services and Applications*, 10(1), 6.

[6] Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., ... & Redmiles, E. M. (2019, January). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 140-149). [CrossRef]

[7] Usman, M. T., Khan, H., Singh, S. K., Lee, M. Y., & Koo, J. (2024). Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices. IEEE Transactions on Consumer Electronics. [CrossRef]

[8] Tardáguila, C., Benevenuto, F., & Ortellado, P. (2018). Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It. *International New York Times*, NA-NA.

[9] Ullah, I., Ali, F., Khan, H., Khan, F., & Bai, X. (2024). Ubiquitous computation in internet of vehicles for human-centric transport systems. *Computers in Human Behavior*, 161, 108394. [CrossRef]

[10] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1-40. [CrossRef]

[11] Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*. [CrossRef]

[12] Rubin, V. L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, 47(1), 1-10. [CrossRef]

[13] Rubin, V. L., Conroy, N., Chen, Y., & Cornwell, S. (2016, June). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7-17).

[14] Gottfried, J., & Shearer, E. (2016). News use across social media platforms 2016.

[15] Campan, A., Cuzzocrea, A., & Truta, T. M. (2017, December). Fighting fake news spread in online social networks: Actual trends and future research directions. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 4453-4457). IEEE. [CrossRef]

[16] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096. [CrossRef]

[17] Chadwick, A., & Vaccari, C. (2019). News sharing on UK social media: Misinformation, disinformation, and correction.

[18] Kogan, S., Moskowitz, T. J., & Niessner, M. (2019). Fake news: Evidence from financial markets. *Available at SSRN*, 3237763.

[19] Zafarani, R., Zhou, X., Shu, K., & Liu, H. (2019, July). Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3207-3208). [CrossRef]

[20] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2), 211-236.

[21] Thompson, S. A. (2017, December 14). *President Trump's lies, the definitive list*. The New York Times - Breaking News, US News, World News and Videos. https://www.nytimes.com/interactive/2017/06/23/opinion/trumps-lies.html

[22] Ur Rehman, I., Ullah, I., Khan, H., Guellil, M. S., Koo,

J., Min, J., ... & Lee, M. Y. (2024). A comprehensive systematic literature review of ML in nanotechnology for sustainable development. *Nanotechnology Reviews*, 13(1), 20240069. [CrossRef]

[23] Graauwmans, V. V. (2016). Fake News in the Online World: An Experimental Study on Credibility Evaluations of Fake News depending on Information Processing Bachelor Thesis Tilburg University.

[24] Guderlei, M., & Aßenmacher, M. (2020, December). Evaluating unsupervised representation learning for detecting stances of fake news. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6339-6349). [CrossRef]

[25] Khan, H., Usman, M. T., Rida, I., & Koo, J. (2024). Attention enhanced machine instinctive vision with human-inspired saliency detection. *Image and Vision Computing*, 152, 105308.

[26] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36. [CrossRef]

[27] Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6), 490-496.

[28] Amorim, E., Cançado, M., & Veloso, A. (2018, June). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 229-237). [CrossRef]

[29] An, J., & Weber, I. (2016). # greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10, No. 1, pp. 523-526). [CrossRef]

[30] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1* (pp. 127-138). Springer International Publishing.

[31] An, J., & Kwak, H. (2017, May). What gets media attention and how media attention evolves over time: large-scale empirical evidence from 196 countries. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 464-467). [CrossRef]

[32] Srivastava, A. (2020). Real time fake news detection using machine learning and NLP. *Int. Res. J. Eng. Technol.(IRJET)*, 7(06).

[33] Khan, H., Jan, Z., Ullah, I., Alwabli, A., Alharbi, F., Habib, S., ... & Koo, J. (2024). A deep dive into AI integration and advanced nanobiosensor technologies for enhanced bacterial infection monitoring. *Nanotechnology Reviews*, 13(1), 20240056. [CrossRef]

[34] Lakshmanarao, A., Swathi, Y., & Kiran, T. S. R. (2019). An effecient fake news detection system using machine learning. *International Journal of Innovative Technology and Exploring Engineering*, 8(10), 3125-3129.

[35] Hiramath, C. K., & Deshpande, G. C. (2019, July). Fake news detection using deep learning techniques. In 2019 1st International Conference on Advances in Information Technology (ICAIT) (pp. 411-415). IEEE. [CrossRef]

[36] Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting fake news using machine learning and deep learning algorithms. In *2019 7th international conference on smart computing & communications (ICSCC)* (pp. 1-5). IEEE. [CrossRef]

[37] Granik, M., & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)* (pp. 900-903). IEEE. [CrossRef]

[38] Ali, D., Iqbal, S., Mehmood, S., Khalil, I., Ullah, I., Khan, H., & Ali, F. (2024). Unleashing the Power of AI in Communication Technology: Advances, Challenges, and Collaborative Prospects. In *Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies* (pp. 211-226). Singapore: Springer Nature Singapore.

[39] Gadekar, P. S. (2019). Fake News Identification using Machine Learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 7(V).

[40] Ivancová, K., Sarnovský, M., & Maslej-Krcšňáková, V. (2021, January). Fake news detection in Slovak language using deep learning techniques. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 000255-000260). IEEE. [CrossRef]

[41] Meesad, P. (2021). Thai fake news detection based on information retrieval, natural language processing and machine learning. *SN Computer Science*, 2(6), 425.

[42] Cai, Y., Pan, S., Wang, X., Chen, H., Cai, X., & Zuo, M. (2020). Measuring distance-based semantic similarity using meronymy and hyponymy relations. *Neural Computing and Applications*, 32, 3521-3534.

[43] Bali, A. P. S., Fernandes, M., Choubey, S., & Goel, M. (2019). Comparative performance of machine learning algorithms for fake news detection. In *Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3* (pp. 420-430). Springer Singapore.

[44] Ali, D., Huque, M. T., Godhuli, J. J., & Ahmed, N. (2022). Detection of Face Emotion and Music Recommendation System using Machine Learning. *International Journal of Research and Innovation in Applied Science*, 7(11), 05-08.

[45] Bangyal, W. H., Qasim, R., Rehman, N. U., Ahmad, Z., Dar, H., Rukhsar, L., ... & Ahmad, J. (2021). Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and mathematical methods in medicine*, 2021(1), 5514220. [CrossRef]

[46] Wu, J., Huang, C., & Chen, Y. (2020, October). Patent Text Classification Study Based on Bi-LSTM-A Model. In *2020 5th international conference on control, Robotics and Cybernetics* (*CRC*) (pp. 1-5). IEEE. [CrossRef]

[47] Ganesh, P., Priya, L., & Nandakumar, R. (2021, June). Fake news detection-a comparative study of advanced ensemble approaches. In *2021 5th International Conference on Trends in Electronics and Informatics* (*ICOEI*) (pp. 1003-1008). IEEE. [CrossRef]

Muhammad Nadeem did hisBachelor's in computer science from GC university Faisalabad Pakistan, and MS.c in Computer Science from Arid Agriculture university Rawalpindi ,Pakistan. Now he is doing his master from Ocean University of China. Reserch work and interest is step length estimation using Neural Network, Machine learning, deep learning and Climate modeling.
(Email: Muhammad.nadeem42300@gmail.com)

Parchamdar Abbas did his Bachelor of Science in Software Engineering from Comsats University Islamabad (CUI), Pakistan and Master of Science in Software Engineering from National University of Modern Languages (NUML), Islamabad, Pakistan. Reserch work and interest is "Natural Language Processing , Machine learning, deep learning and AI. (Email: pdabbas137@gmail.com)

Wei Zhang is associate professor and his reserch work and interest is step length estimation using Deep Learning, Machine Learning, Climate modeling, Multi-task Learning. (Email: Weizhan@ouc.edu.cn)

Sumaira Rafique received the MS degree in Data Science from National University of Computer and Emerging Sciences, Lahore, Pakistan in 2024. Her research interests include Natural Language Processing, Generative AI, Deep Learning, Machine Learning, Image Processing. Currently she is working as a Machine Learning Engineer at Reallytics.ai, Lahore. (Email: rsumaira80@gmail.com)

Sundas Iqbal received the MS degree in Computer Science from National college of Business administration and Economics, Pakistan in 2018. Her research interests include Biomedical image analysis, Machine Learning, Image Processing. Currently she is PhD student at Nanjing University of Information Science and Technology (NUIST) with a research focus on domain adaptation, transfer learning and novel AI techniques for various fields. (Email: Sundasiqbal058@gmail.com)